arXiv:2301.01811v1 [stat.ME] 4 Jan 2023

利用可复现希尔伯特空间分析重复空间点过程。

A. Sim´o数学系-IMAC。海梅一世大学。Riu Sec 大道，无门牌号。12071-卡斯特利翁，西班牙。

2023年1月6日

摘要

本文重点探讨再生核希尔伯特空间理论在重复点过程统计分析中的应用。我们表明，空间点过程可以视为再生核希尔伯特空间中的随机变量，因此，适用于这些空间统计分析的方法论和理论结果也可应用于空间点过程。具体而言，作为示例，我们展示如何利用所提出的方法论，通过 MBox 和多元方差分析检验识别若干类重复点模式之间的差异，并使用判别函数对新的观测进行分类。

关键词

再生核希尔伯特空间；函数型数据分析；重复空间点过程；方差分析；监督分类

1 引言

空间点过程是一种随机过程，其实现是在研究区域 E of $\mathbb{R}^2$ 中局部有限的点集（事件的位置）。局部有限是指，对于任意有界博雷尔集，其中的点数有限（概率为 1）。空间点模式是许多问题中自然产生的采样信息。例如，森林中的树木位置、天空中的星系、城市中的某类商业设施，或某种疾病的病例分布。关于点过程理论及其应用的经典著作包括 Stoyan等（1988）；Stoyan和 Stoyan（1994）；Baddeley等（2007）；Illian等（2008）；Diggle（2013）；Cressie（2015）。

用数学术语来说，如果 $(E)^n \equiv E \bullet E \bullet \cdots \bullet E\,(n \text{ times})$ 是 $n$elements of E 的集合；

我们将指数空间定义为

$$E_e \equiv \bigcup_{n=0}^{\infty} (E)^n.$$

点过程 Φ 被定义为从概率空间 $(\Omega\ \mathcal{A}, \mathcal{P}')$ 到可测空间 $(E_e, \mathcal{B}_e)$ 的可测映射。有关可测指数空间的详细信息，请参阅 Carter 和 Prenter（1972）。

文献中广泛采用准确且有充分理论依据的点过程分析方法，但其中大多数方法都聚焦于只能获得过程一次观测的情形。

表示和／或描述点过程主要有两种不同的方法：事件密度与分布，以及随机计数测度。本文聚焦于第二种方法。我们的目标是利用随机测度与取值于再生核希尔伯特空间的随机变量之间的关系，处理由再生核希尔伯特空间中的随机函数刻画的点过程。因此，我们探讨如何在这一空间中采用更复杂的统计方法。具体而言，我们提出的方法能够非常自然且直接地分析重复点模式，即由多个点模式组成的数据集，其中这些点模式可视为同一实验的独立重复样本。

再生核希尔伯特空间在统计学文献中广为人知，并常用于机器学习领域（Berlinet and Thomas-Agnan, 2011；Saitoh and Sawano, 2016），作为解决欧几里得空间或 L^2 空间，甚至黎曼流形上的分类或回归问题的有力工具。特别是在欧几里得空间中，许多分类算法的成功得益于核方法的应用（Sch¨olkopf et al., 2002）。然而，当数据取值于再生核希尔伯特空间时，关于统计方法的文献几乎空白（Luki´c and Beder, 2001）；而这正是本文所讨论的情形。

函数型数据统计理论是统计学中的一个重要研究领域。它研究的是这样一种样本：对每个个体都观测到一个函数。Silverman 和 Ramsay（2005）；Ferraty 和 Vieu（2006）；Horv´ath 和 Kokoszka（2012）以及 Aneiros 等人（2017）的著作都是重要参考文献，Cuevas（2014）以及 Goia 和 Vieu（2016）的精彩综述也同样值得参考。尽管函数型数据分析理论已纳入经典参数统计或非参数统计中的许多工具，但样本空间的无限维特性仍带来了特殊问题（Ferraty 和 Vieu，2006），即便在实践中，我们只需将观测到的曲线采样为有限个观测点。

关于函数型数据有两种不同的观点。第一种观点认为，函数型数据是取值于希尔伯特空间的随机变量的实现。第二种观点则认为，函数型数据是具有光滑均值函数和协方差函数的随机过程的样本轨迹。从理论角度看，这两种观点之间存在微妙差异。

再生核希尔伯特空间常见于函数型数据分析的第二种视角中（Preda, 2007；Kadri 等，2016），因为 Lo´eve-Parzen 一致性（Aronszajn, 1950）将二阶随机过程与由其协方差函数生成的再生核希尔伯特空间联系起来（Eubank 和 Hsing, 2008；Kupresanin 等，2010）。函数型主成分方向最终构成与协方差相关的希尔伯特-施密特协方差算子的一个正交归一基

核（Horv´ath 和 Kokoszka, 2012；Cuevas, 2014）。不过，在我们的情形中，函数型数据按定义是希尔伯特空间中一个变量的样本（第一种观点），我们的近似方法将与后一种观点类似。

在 Baddeley（2015）中，作者借助 R 软件包 $Spatstat$ 介绍了重复点模式的实际分析。本文将使用该软件包中的一个数据集，即锥体数据集。该数据集包含 Diggle 等人（1991）的数据，记录了 12 名正常、9 名患有分裂情感性障碍以及 10 名患有精神分裂症的人类受试者脑中锥体神经元的位置。我们的所有实现均使用 R 编写（R 核心团队，2021年），主要使用 $Spatstat$ 和 $MASS$ 软件包。

本文的结构安排如下:
首先，第2节介绍理论概念。其次，第3节将这些理论概念应用于重复点模式的统计分析。之后，第4节详述两个具体应用。最后，第5节讨论结论。

2 从点过程到再生核希尔伯特空间中的随机元素

如引言所述，本文旨在展示一种方法，通过在再生核希尔伯特空间（RKHS）中利用随机函数刻画点过程，从而对其进行研究。本节将介绍实现这一目标所必需的理论概念。首先，简要介绍再生核希尔伯特空间的定义及其性质。其次，说明如何将测度嵌入再生核希尔伯特空间。随后，将点过程表示为随机测度，并讨论希尔伯特空间中随机变量的一般理论结果，尤其是再生核希尔伯特空间中的相关结果。

2.1 再生核希尔伯特空间

再生核希尔伯特空间（RKHS）是一个函数希尔伯特空间$f:\ E\ \rightarrow\ \mathbb{R}$，具有一些实用且有趣的性质。再生核希尔伯特空间理论由 Aronszajn（1950）建立。

定义 1。设 H 是定义在 E上的实值函数希尔伯特空间，且 $\langle\cdot,\cdot\rangle_H$ 是 H上的内积。若函数 $k:\ E\times\mathbb{R}^n\rightarrow\mathbb{R}$满足以下条件，则称其为与 H 相关联的再生核（*rk*）：

1. 对于每个 $x\in E$, $k(,x)\ \in H$。
2. k 满足“再生性质”；也就是说，$\forall f\in H$ 和 $x\in E$

$$f(x)=\langle k(\cdot,x),f\rangle_H$$

定义 2。一个实值函数希尔伯特空间若与一个再生核（*rk*）相关联，则称为再生核希尔伯特空间。

再生核希尔伯特空间可以由一个核构造得到，而每个核都唯一确定一个再生核希尔伯特空间（Moore-Aronszajn 定理），记为 H_k。H_k 的构造如下。

我们考虑以下线性组合的集合： H_0

$$H_0 = \{\sum_{i=1}^{N} b_i k(\cdot, x_i), N \geq 1, b_i \in \mathbb{R},\ x_i \in E\},$$

与核 k 相关联的再生核希尔伯特空间 H_k 是 H_0的闭包。

给定 $\phi_1 = \sum_{i=1}^{N_1} a_i k(\cdot, x_i)$ 和 $\phi_2 = \sum_{i=1}^{N_2} b_i k(\cdot, y_i)$，$\langle \phi_1, \phi_2 \rangle_k = \sum_i \sum_j a_i b_j k(x_i, y_j)$

2.2 将测度嵌入再生核希尔伯特空间

研究随机测度需要精妙的数学工具。因此，遵循 Berlinet and Thomas-Agnan（2011）的做法，我们首先说明如何利用再生核将测度表示在函数空间中，并从狄拉克测度入手。随后，我们将利用这种嵌入来定义内积。

给定 $\mathbb{R}^2$ 的一个紧子集 E，以及 E 子集的博雷尔代数 $\sigma-\ \mathfrak{B}$，狄拉克测度 δ_x 对于 E 中的 x 定义如下：

$$\delta_x(A) = \begin{cases} 1 & \text{if} \quad x \in A \\ 0 & \text{if} \quad x \notin A \end{cases}$$

其中，A 是 E中的一个博雷尔集。

映射：

$$\delta_x \mapsto k(\cdot, x)$$

将 E 上的狄拉克测度集合嵌入到核为 k的再生核希尔伯特空间中。

如果函数 $k(y, \cdot)$是可测的，那么函数 $k(\cdot, x)$在点 y 处的值 $k(y, x)$可以写成积分 $\int k(y, t) d\delta_x(t)$，于是该映射可以改写为：

$$\delta_x \mapsto \mathcal{I}_{\delta_x} = \int k(y, t) d\delta_x(t)$$

此外，对于 H_k $\langle f, k(\cdot, x) \rangle_H = f(x)$ $= \int f(t) d\delta_x(t)$中的任意可测函数 f

更一般地，如果 $x_1, ..., x_N$是 E中的 N个互不相同的点，且 $b_1, ..., b_N$、N是非零实数，那么线性组合

$$\mu = \sum_{i=1}^{N} b_i \delta_{x_i}$$

狄拉克测度的有限线性组合称为有限支撑符号测度。我们可以用以下内容扩展之前的映射

$$\mu \mapsto \sum_{i=1}^{N} b_i k(\cdot, x_i) = \int k(y, t) d\mu(t). \tag{1}$$

该映射将具有有限支撑的 E 上的测度集合 H_k $\mathcal{M}_0$ 嵌入其中，而集合 H_0 可视为其在 H_k 中的表示元集合。同样，我们具有如下性质：对于 H_k $\langle f, \sum_{i=1}^{N} b_{ik}(\cdot, \ x_i)\rangle_H = \sum_{i=1}^{N} f(x_i) = \int f d\mu$. 中的任意可测函数 f

根据上述推广，我们还可以通过以下映射将 $\mathcal{M}$ 上的符号测度集合嵌入 H_k：

$$\begin{aligned} \mathcal{M} &\to H_k \\ \mu &\mapsto \mathcal{I}_\mu = \int k(\cdot, t) d\mu(t) \end{aligned}$$

更多细节请参见 Berlinet and Thomas-Agnan（2011）。

如果我们假设 k 满足：当 μ 和 ν 不相等时，函数 I_μ 和 I_ν 不同，那么 Berlinet and Thomas-Agnan （ 2011 ）的定理 99 保证以下映射：

$$\begin{aligned} \mathcal{M} \times \mathcal{M} &\to \mathbb{R} \\ (\mu, \nu) &\mapsto \langle \mathcal{I}_\mu, \mathcal{I}_\nu \rangle_H \end{aligned}$$

在 $\mathcal{M}$ 上定义了一个内积，使得 $\mathcal{M}_0$ 在 $\mathcal{M}$ 中稠密，反之亦然。

Guilbart（1979）率先研究了再生核与空间 $\mathcal{M}$ 上内积之间的关系。他利用这一嵌入，刻画了在测度集合上诱导弱拓扑的内积。

2.3 点过程与随机测度

若 Φ 是一个点过程，则与 Φ 相关联的随机计数测度定义如下：

$$\mu_\Phi(B) = \#\{x : \ x \in \Phi \cap B\}, \tag{2}$$

其中 B 是 E中的一个博雷尔集。

点过程的随机计数测度刻画了其概率分布，并为将点过程理论纳入随机测度的一般理论提供了框架（Daley 和 Vere-Jones，2008）。

若 $\mathcal{M}$ 是 E 中的一组测度，并配备了某个 σ-代数，则随机测度可视为一个取值于 $\mathcal{M}$的随机变量。关于随机测度理论的详细研究，参见 Daley 和 Vere-Jones（1998）。尽管这一概念看似简单，但随机测度的定义本身却引出了微妙的问题。例如，由 $\mathcal{M}$ 上的某种拓扑可能导出的 σ-代数并非易事，而由此产生的理论也涉及微妙的数学问题。

因此，在了解如何将测度集合嵌入再生核希尔伯特空间之后，我们利用这一嵌入，将随机测度定义并视为再生核希尔伯特空间中的随机元素，即取值于再生核希尔伯特空间的随机变量。我们假设

该随机变量以概率 1 取值于 $\mathcal{M}$ (或 $\mathcal{M}_0$)。这是 Suquet，1986 提出的构造。

通过这一构造，我们随后可以利用关于可分希尔伯特空间中概率律的已知结果。此外，再生核希尔伯特空间是向量度量空间，可以视为通常欧几里得空间的自然延伸。在欧几里得空间中建立的大多数理论和方法统计结果，都可以直接延伸到再生核希尔伯特空间中。此外，希尔伯特空间的完备性提供了一个框架，使我们能够将无限维向量视为有限维向量的极限。

巴拿赫空间和希尔伯特空间中的概率论是现代概率论的一个重要分支。关于这一主题的完整论述，可参见 Ledoux 和 Talagrand（1991）以及 Hsing 和 Eubank（2015）。许多关于大样本结果的结论都可应用于可分希尔伯特空间。在这些理论结果中，我们在此仅回顾一个中心极限定理，该定理保证了收敛到与欧几里得空间中类似的高斯过程。完整论述请参见 Hsing 和 Eubank（2015）。

定理 3. 设 $\chi_1, ..., \chi_n$ 是希尔伯特空间 H 中独立同分布的随机元素，其均值为 *0*，且 $\mathbb{E}\|\chi_i\|^2 < \infty$。则

$$\xi_n := \frac{\sum_{i=1}^n \chi_i}{\sqrt(n)} \to^d \xi,$$

其中，ξ 是 H 中的高斯随机元素，其协方差算子等于 $\mathbb{E}(\chi_i \otimes \chi_i)$。$\otimes$ 是 H 中的张量积算子，$\|\cdot\|$ 是由 H 中的内积定义的范数，而 $\to^d$ 表示依分布收敛。

此外，如果该希尔伯特空间还是一个再生核希尔伯特空间，我们就能得到更多重要且强有力的性质，从而保证标准统计技术可应用于再生核希尔伯特空间中的随机变量。Guilbart（1979）证明了一个 Glivenko-Cantelli 定理，并将其应用于估计和假设检验。Berlinet（1980b,a）研究了再生核希尔伯特空间上的概率集合中的弱收敛，以及再生核希尔伯特空间值变量的可测性和可积性。另一个重要结果（尽管本文不会用到）是《Hsing和Eubank》（2015）的定理 7.5.1，该定理指出，再生核希尔伯特空间中的随机变量也是一个随机过程，反之亦然。

转向点过程的情形，方程 2 可改写为：

$$\mu_\Phi(B) = \sum_{x_i \in \Phi} \delta_{x_i}(B) = \sum_i^N \delta_{x_i}(B), \tag{3}$$

其中 B 是 E 中的一个博雷尔集，δ 表示狄拉克测度。

由 Φ 定义的计数测度是一种随机测度，前面提到的关于随机测度的所有内容均适用于重复点过程的统计分析。

计数测度是有限支撑测度的一种特殊情况，我们可以通过映射 1 将其嵌入再生核希尔伯特空间中：

$$\mu_\Phi \mapsto \sum_{i=1}^{N} k(\cdot, x_i). \tag{4}$$

此外，我们假设其对应的再生核希尔伯特空间随机变量以概率 1 取值于 E、$\mathcal{N} \subset \mathcal{M}_0$ 上的计数测度集合。

此外，满足以下条件：如果两个计数测度 μ和 ν不相等，即它们的支撑集 $\{x_1, ..., x_N\}$和 $\{y_1, ..., y_M\}$不同，那么 $I_\mu = \sum_{i=1}^{N} k(\cdot,\ x_i)$ 与 $I_\nu = \sum_{i=1}^{M} k(\cdot,\ y_i)$ 也不同。

在计算方面，映射 4 可以通过例如 R 的 spatstat程序包中的函数 密度 轻松获得。

尽管这种工作方式带来了巨大的可能性，但本文仅聚焦于两个应用，以作说明。

在第一个应用中，我们有不同的实验组，在每个组内观察一个点过程的独立重复样本，并希望检验组间是否存在差异；换言之，这是一个响应变量为点模式的方差分析问题。在第二个例子中，我们同样有不同的组，但我们希望找到一条规则，将新的点模式分类到其中一个组中；换言之，这是一个解释变量为点模式的监督分类问题。

在这两种情况下，都将采用经典多元统计方法，以充分利用前文提到的所有理论结果。

3 应用于复制点过程的统计分析

在上一节中，我们已经看到如何将点过程转换为再生核希尔伯特空间中的随机元素；因此，在本节中，我们假设有一个来自再生核希尔伯特空间 H_k 中随机元素的样本 $\{\varphi_i(\cdot)\}_{i=1}^{n}$，该样本中的每个元素具有如下表达式：

$$\varphi_i(x) = \sum_{l=1}^{N_i} k(x, x_{il}), x \in E \tag{5}$$

由于我们的数据是一类特殊的函数，因此利用函数型数据分析（FDA）的见解来开展统计分析是合理的。不过，我们的数据与函数型数据分析文献中的典型数据截然不同。首先，我们的数据并未采用标准形式表示；在标准形式下，第 i 个函数型数据仅由一组离散测量值 $((y_{i1},\ t_1) \cdots, (y_{im_i},\ t_{m_i}))$ 给出，而我们对函数的解析形式知之甚少。样本中的每个数据都是一个函数 $\varphi_i(x)$，其解析表达式由（式 5）定义。

其次，我们的函数空间是一个再生核希尔伯特空间（RKHS），也就是说，我们的原始函数型数据“存在于”再生核希尔伯特空间中；如前所述，其向量结构和内积可用于数据分析。因此，我们提议将数据集中的每个函数表示为由定义该再生核希尔伯特空间的核的特征函数分解所给出的标准正交基上的展开。核的特征函数分解所给出的各元素彼此正交，并按照最优逼近准则排序。这些性质使我们能够降低维数，从而像在多元欧几里得情形中一样应用经典统计程序。此前，研究人员曾在截然不同的背景下使用过类似的思想：对几何对象进行监督分类。以下结果与 Barahona 等人（2018）第3节的结果类似。

为了得到在这个基底上的投影，我们首先需要改写函数的表达式，使得在样本的所有点模式中，核的求值点都相同。为此，我们可以使用以下定理。

定理 4（表示定理）。给定 $\varphi_i(\cdot)$、网格 $\{a_l\}_{l=1}^N$ *in* E，记作$\varphi_i(a_l) = b_{il}$，并给定正则化参数 $\gamma > 0$，则 $\exists! \chi_i\colon \mathbb{R}^2 \to \mathbb{R}; \chi_i(y) = \sum_{l=1}^N \beta_{il} k(y, a_l)$ 满足：

$$\chi_i = \underset{g \in H_k}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} (g(a_l) - b_{il})^2 + \gamma \left\| g \right\|_{H_k}^2, \tag{6}$$

其中 $\beta_{il} \in \mathbb{R}$（对于 $l = 1, \ldots, N$）是以下方程的解：

$$(\gamma\, N\, \mathbb{I}_{N \times N} + K|_a)\beta_i = b_i,$$

其中 $K|_a$ 是定义为 $(K|_a)(i, j) = k(a_i, a_j)$ 的矩阵，$i, j = 1, \ldots, N$ 和 β_i, b_i 是向量

$$\beta_i = \begin{pmatrix} \beta_{i1} \\ \beta_{i2} \\ \vdots \\ \beta_{iN} \end{pmatrix}; \quad b_i = \begin{pmatrix} b_{i1} \\ b_{i2} \\ \vdots \\ b_{iN} \end{pmatrix}$$

应用该定理后，从现在起我们将使用以下样本：

$$\{\chi_i(\cdot) = \sum_{l=1}^{N} \beta_{il} k(\cdot, a_l)\}_{i=1}^n \tag{7}$$

众所周知（Hsing和Eubank，2015），如果 E 是紧致的，且 k 连续、可测且有界，那么与核函数 k 相关联并由下式定义的积分算子：

$$(\mathcal{K}f)(\cdot) = \int_E k(\cdot, x) f(x) dx, \quad f \in \mathcal{L}_2(E), \tag{8}$$

其中 $\mathcal{L}_2(E)$ 是 E 上平方可积函数所构成的空间，是一个紧致、连续、自伴且为正算子的算子。

因此，它可以表示为

$$\mathcal{K} = \sum_{q \geq 1} \lambda_q e_q \otimes e_q,$$

其中，$\{\lambda_q, e_q\}_q$ 是其特征值和（正交归一）特征函数的可数序列（谱分解）。

此外:

$$k(x, y) = \sum_{q \geq 1} \lambda_q e_q(x) e_q(y)$$

.

并且，$\{\sqrt{\lambda_q} e_q\}_q$ 是 H_k的一组完备正交归一基。

以下结果说明了如何将样本投影到再生核希尔伯特空间的这组基底上。

定理5. 设 χ_i 是再生核希尔伯特空间 H_k中的一个元素，它可以表示为

$$\chi_i(\cdot) = \sum_{q=1}^{\infty} \mu_{qi} \left(\sqrt{\lambda_q} e_q(\cdot) \right), \tag{9}$$

其中 $\{\sqrt{\lambda_q e_q}\}_{q=1}^{\infty}$。

In addition, if $\chi_i(\cdot) = \sum_{q=1}^{\infty} \mu_{qi} \left(\sqrt{\lambda_q} e_q(\cdot) \right)$ *and* $\chi_j(\cdot) = \sum_{q=1}^{\infty} \mu_{qj} \left(\sqrt{\lambda_q} e_q(\cdot) \right)$,

内积为: d

$$\langle \varphi_i, \varphi_j \rangle_k = \sum_q \mu_{qi} \mu_{qj}.$$

此外（*Smale* 和 *Zhou*，*2009*年；*Gonz ´alez* 和 *Mu ˜noz*，*2010*），前$d =$ 秩 $(K|_a)$ 个系数 μ_{qi} 可以近似为

$$\widehat{\mu_{qi}} = \sqrt{\ell_q} (v_q \cdot \beta_i) \tag{10}$$

其中 $v_q \in \mathbb{R}^N$ 是 $K|_a$的特征向量，ℓ_q 是 $K|_a$的特征值。

以下结果表明，如果我们用这个基表示无限维数据，并将其截断以获得有限维向量样本，就能达到最高精度。

命题6. *Hsing* 和 *Eubank*（*2015*）中的定理*4.4.7*和*4.6.8*告诉我们，对于固定整数 $r > 0$ 且满足 $\lambda_r > 0$:

$$\min_{f_1, \ldots, f_r \in H_k} \int \int_{E \times E} \left(k(y, x) - \sum_{q=1}^{r} f_q(y) f_q(x) \right)^2 dy dx,$$

最小值由 $f_q = e_q$ 达到。

这一结果确保，截断特征值-特征向量分解能为 k 提供最佳近似，从而为我们的数据提供最佳近似（式 7）。

如果我们将公式 9 中的求和截断为较少的项 $rr \leq d = 秩(K|_a)$，则每个点模式 φ_i 对于 $i = 1, \ldots\ldots, n$，可由正交归一基$\{\sqrt{\lambda_q}\psi_q\}_{q=1}^{\infty}$上的系数 μ_{qi} 对于$q = 1, \ldots\ldots, r$（由 $\widehat{\mu_{qi}}$估计）给出。因此，它可以表示为 r维向量

$$\mu_i = (\widehat{\mu_{1i}}, \widehat{\mu_{2i}}, \ldots, \widehat{\mu_{ri}}). \tag{11}$$

这一表达式将无限维问题最优地化简为有限维问题。现在可以应用众所周知的经典多元方法。具体而言，在下一节中将使用两种众所周知的经典多元方法进行说明：多元方差分析和判别分析。

4 应用于人脑中锥体神经元的位置

如前所述，锥体数据集 包含在 R 软件包 $spastat$中，数据来自 Diggle 等（1991）。这些数据是人脑中锥体神经元的位置。31 名人类受试者中的每一人均观测到一个点模式。其中有 12 例正常对照、9 例分裂情感性障碍和 10 例精神分裂症病例。关于这些数据的更详细说明，请参见 Diggle 等（1991）。图 1给出了我们样本中各点模式的绘图，每位受试者对应一幅图。

在此示例中，使用 $E = [0,1] \times [0,1]$ 和参数为 $\sigma = 0.05$ 的高斯核，通过映射 4 将点模式嵌入再生核希尔伯特空间。使用网格 $\{a_l\}_{l=1}^{N}$，其在 E 内包含 a_l 个等距点，垂直和水平步长均为 $h = 0.02$，以应用表示定理 4。方程 6 中的参数 $\gamma = 0.000127$ 固定为使矩阵 $(\gamma\, N\, \mathbb{I}_{N\times N} + K|_a)$ 可逆的最小值。参数 h 和 σ 与工作数据的精度有关。当 h 和 σ 减小时，精度会提高，但计算成本和计算复杂度也会增加。所选数值是在两者之间取得的平衡。

应用表示定理后，我们得到一个样本 $\{\chi_i(\cdot)\ y_i\}_{i=1}^{31}$，其中包含$\chi_i(\cdot) = \sum_{l=1}^{N} \beta_{i,l} k(\cdot, a_l)$，以及 y_i 一个分类变量，用于指示受试者属于正常、分裂情感性障碍还是精神分裂症。

在图2中，我们可以看到每组中一个点模式的图，以及它在再生核希尔伯特空间中的对应元素 φ_i 和 χ_i。

最后，应用公式 9，得到系数向量 μ_i，它表示各点模式相对于 H_k的标准正交基的表示。由于样本量非常小（每组分别有 12、9 和 10 个案例），我们在 $r = 6$处截断。

分析基底中前几个函数的含义很有必要。在图3中，我们可以看到前六个特征函数。第一个函数的系数

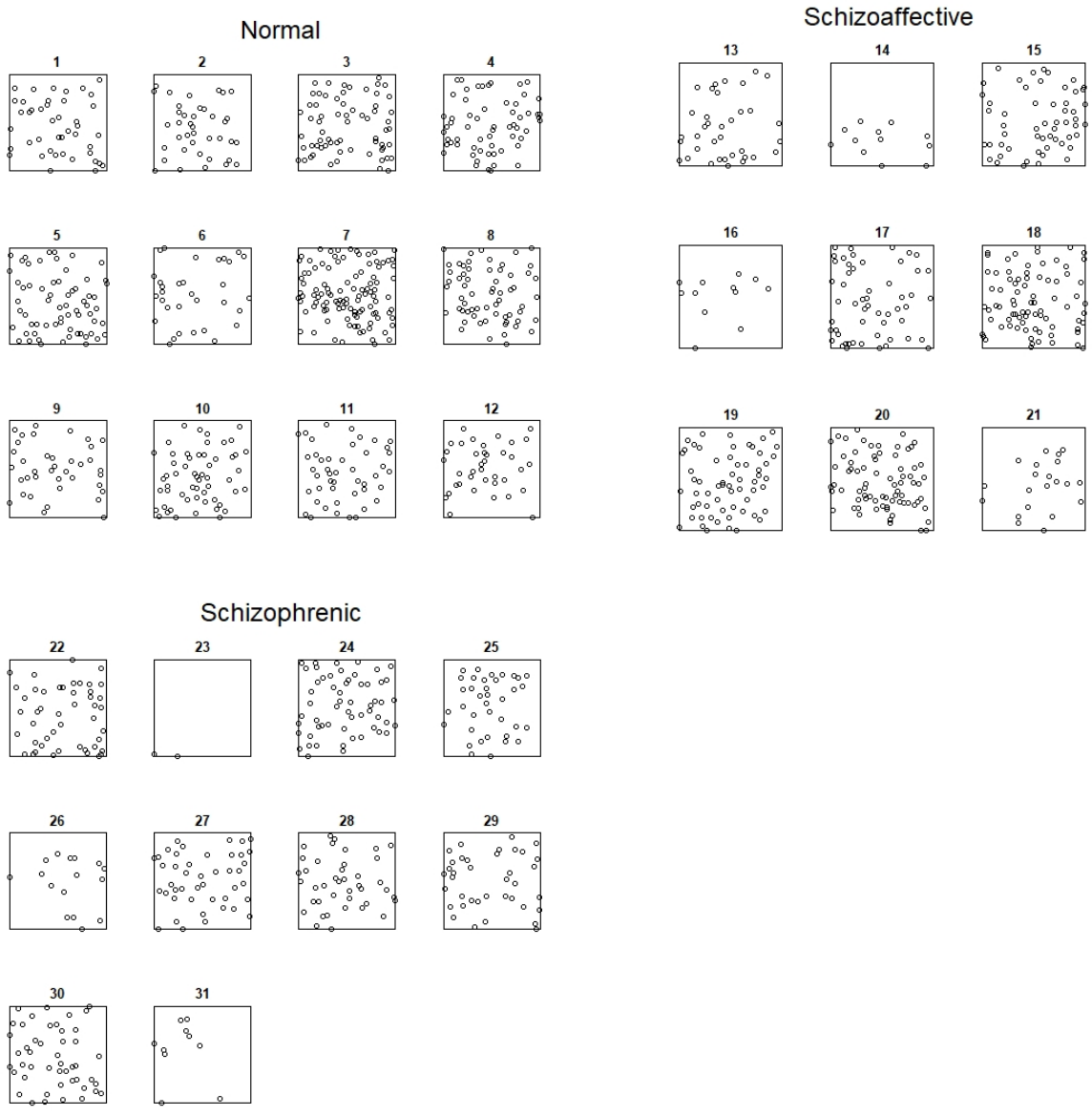


图1：三组受试者的锥体神经元位置。

函数测量窗口中心区域的高密度，第二个和第三个函数测量窗口一半区域内的高密度，以此类推。

4.1 方差分析

空间点模式文献中有两种主要方法，可用于识别多个实验组之间的显著差异（Diggle 等人，2000）。第一种方法（Diggle 等人，1991；Ram´on 等人，2016；Gonz´alez Monsalve 等人，2018）基于模式的函数描述符和非参数推断。

第二种方法基于对模式采用参数模型的假设（通常为成对相互作用点过程或吉布斯过程），针对每个组，使用最大似然或伪似然方法估计模型参数，并通过比较假设参数相同与不作此假设时的拟合结果来检验组间差异（Illian 和 Hendrichsen，2010；Illian等，2012）。

按照上述步骤，我们将使用嵌入得到的再生核希尔伯特空间元素。

由于再生核希尔伯特空间是一个向量空间，因此样本均值元素可直接计算为：$\bar{\chi}(\cdot) = \sum_{l=1}^{N} \sum_{i}^{n-\beta_i} {}_{n}{}^{l}\, k(\cdot, a_l)$。图4展示了各组的样本均值元素。

在本例中，我们只关注以下问题的答案：各组观测到的模式在均值上是否存在显著差异？为此，可以对 r元样本$\{\mu_i, y_i\}_{i=1}^{31}$进行多元方差分析检验。在进行多元方差分析检验之前，必须先进行Box M检验，以检验方差-协方差矩阵齐性这一假设。两项检验的结果见表1。在显著性水平为0.05时，方差-协方差矩阵齐性的假设将被接受，而均值相等的假设将被拒绝。

需要注意的是，在这种情况下，样本量较小，因此中心极限定理不适用于假定数据服从正态分布。对残差进行 Shapiro 单变量检验，结果均不显著。尽管这些结果不能保证多元正态性，但有两个原因表明无需担忧。首先，多元方差分析检验对非正态性的稳健性较好；其次，本例仅用于说明。

拒绝均值相等这一多元假设后，我们进行单因素方差分析检验，以更详细地分析差异。基底前六个系数对应的 p 值分别为：0.069、0.044、0.066、0.42、0.58 和 0.115。由此可以得出，类别间的差异主要体现在图3中前四个函数所对应的四个系数上。

再次强调，本节的目标是展示所提出的方法论的潜在应用。若要得出临床结论，则有必要与神经解剖学专家合作开展更深入的研究。

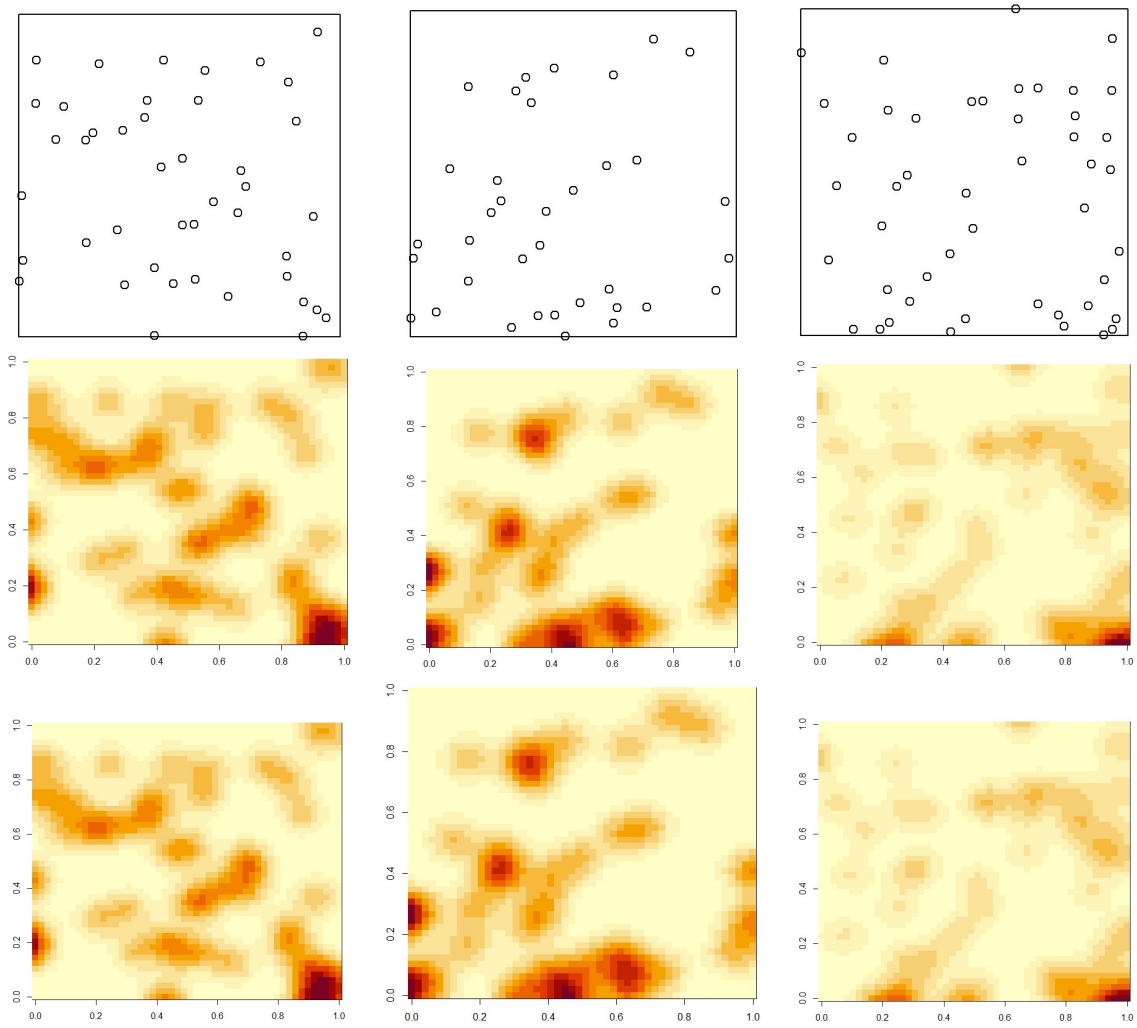


图2：应用表示定理前后各类别第一位受试者的点模式和RKHS函数。

Test	统计量	df	p值
Box的M检验	36.009	42	0.7304
多元方差分析	2.2358	12-48	0.02451

表1：Box的M统计量和多元方差分析结果。

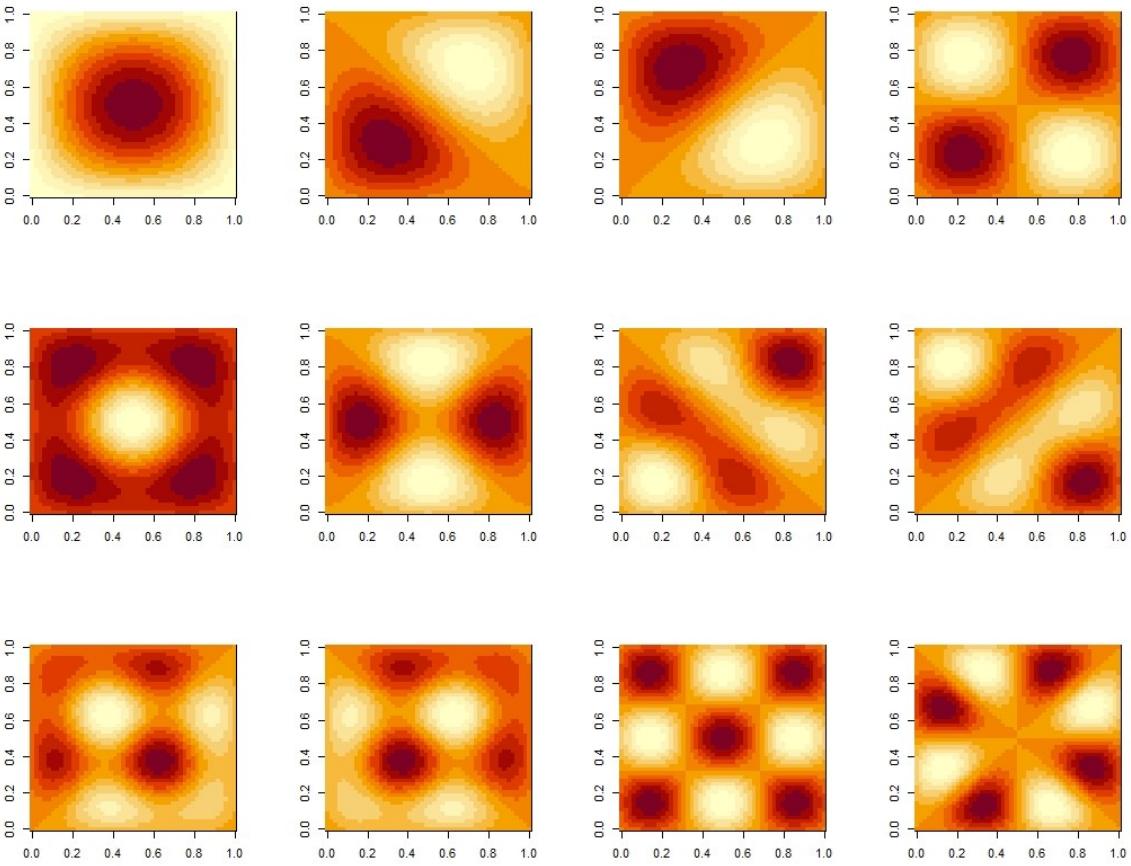


图3： $[0,1]\times[0,1]$ 窗口中高斯核的前几个特征函数。

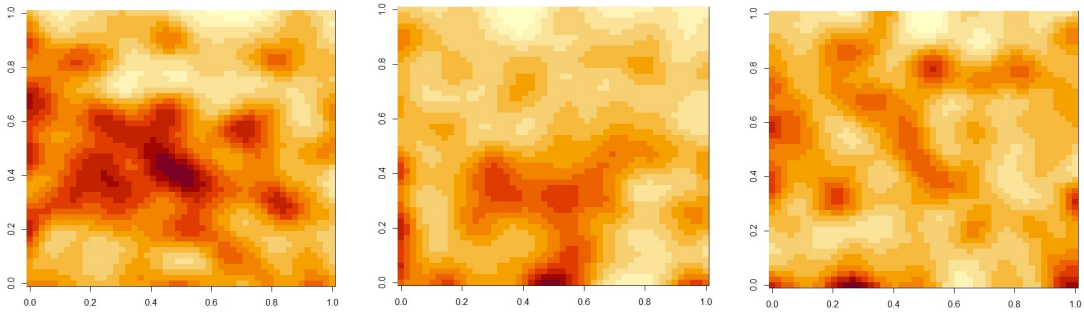


图 4：各类别的 RKHS函数均值：正常、分裂情感性障碍和精神分裂症。

4.2 判别分析

发现受试者组之间存在显著差异后，根据其空间模式制定判别规则，以将新个体分类为正常、分裂情感性障碍或精神分裂症患者，会非常有用。

据我们所知，将监督分类方法应用于重复点模式的相关文献较少，且主要依赖基于不相似度的方法（Mateu 等人，2015；Pawlasov´a 和 Dvo˘r´ak，2022）。Gallego 等人（2016）也曾使用类似方法对胚芽-颗粒模型进行监督分类。

众所周知，关于监督分类方法的文献十分丰富，涵盖了大量方法，从基于多元统计学的方法到最新的深度学习技术（Hastie 等人，2020）。但是，如前所述，本文仅关注经典多元统计方法，因为定理 3使我们能够利用概率参数模型和收敛定理。因此，我们将使用线性和二次判别函数（Venables 和 Ripley，2013）来说明本文提出的新方法。众所周知，这两种方法都是参数贝叶斯方法，假设解释变量服从多元高斯模型，并分别假设方差-协方差矩阵相等和不相等。每个类别的先验 概率将根据样本进行估计。

由于我们的数据集非常有限，我们将首先通过模拟点模式展示一些示例。我们进行了两项不同的实验。

在第一个实验中，在一个面积为 1 的矩形窗口内，模拟了两个样本，每个样本包含 20 个来自强度分别为 λ_1 和 λ_2 的齐次泊松点过程（HPPP）的观测。该实验又重复了两次，强度分别为 $\lambda_1 = 50$ 和 $\lambda_1 = 90$，且两种情况下均为 $\lambda_2 = 100$ 。图 5 展示了每个类别的一个点模式及其在再生核希尔伯特空间中的对应元素。由于这两类点过程的差异体现在其

计数测度的均值上，因此两种情况下都使用了线性判别函数（R 包 线性判别分析 *MASS* 中的函数）。表 2 展示了训练误差和交叉验证误差；即使在第二种情况下，两种模型之间的差异非常微小，也取得了出色的结果。

在第二个实验中，再次在 $[0,1] \times [0,1]$ 窗口内模拟了两个不同点过程的两个大小为 30 的样本。第一个样本对应于强度为 $\lambda_1 = 36$ 的齐次泊松点过程，第二个样本对应于中心泊松过程强度为 $\kappa = 6$ 的泊松簇点过程（PCPP），其中每个簇由圆盘内的 6 个点组成，圆盘半径为 0.2。由此得到的强度也为 $\lambda_2 = 36$。图 5 再次展示了每个类别的一个点模式及其在再生核希尔伯特空间中的对应元素。在这种情况下，两个点过程具有相同的强度，两个类别之间的差异体现在空间变异性上。因此，线性判别函数效果不佳，必须使用二次判别函数（R 包 MASS 中的 二次判别分析 函数）。

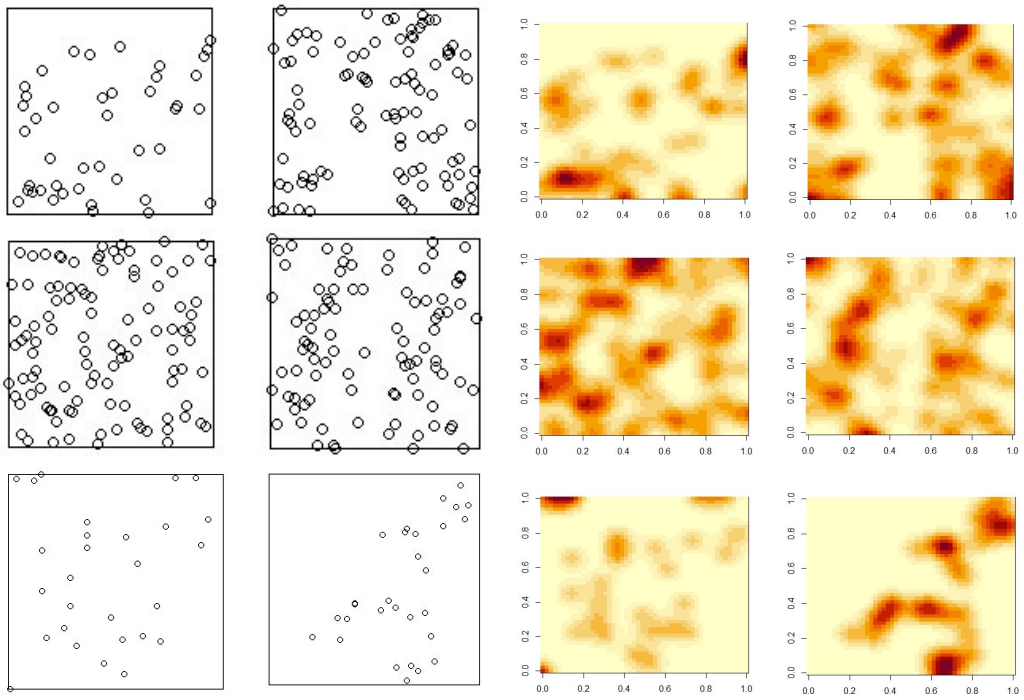


图 5：每个类别的一个模拟点模式及其对应的再生核希尔伯特空间元素。第一行：参数为 $\lambda_1 = 50$ 和 $\lambda_2 = 100$ 的齐次泊松点过程；第二行：参数为 $\lambda_1 = 90$ 和 $\lambda_2 = 100$ 的齐次泊松点过程。第三行：参数为 $\lambda_1 = 36$ 的齐次泊松点过程，以及参数为 $\lambda_2 = 36$ 的泊松簇点过程。

采用了上一节中相同的准确度参数，即 $\sigma = 0.02$ 和 $h = 0.05$。对于变量数量，即 r 的值，测试了从 r=6 到 r=10 的不同取值，但未发现明显差异。总体而言，r=7 时结果最佳，相关结果列于表 2。结果再次十分出色。除了交叉验证误差极小之外，分类正确样本的事后 概率实际上都大于 0.95，而错误分类样本的概率则低于 0.6。

模型	强度	训练误差	交叉验证误差
HPPP-HPPP	$\lambda_1 = 50, \lambda_2 = 100$	0	0
HPPP-HPPP	$\lambda_1 = 90, \lambda_2 = 100$	0.1	0.1755
HPPP-PCPP	$\lambda_1 = 36, \lambda_2 = 36$	0.05	0.117

表 2：模拟实验中的监督分类误差。

在测试所提出的方法论在监督分类问题上的表现后，我们将线性判别分析函数应用于实际案例，以制定判别规则，将新个体分类为正常、分裂情感性障碍或精神分裂症患者。我们采用线性判别分析，因为上一节用于检验方差齐性假设的 Box M 检验未发现显著差异。正如预期，由于样本的局限性，结果并不理想。训练误差相对较小（0.29），但交叉验证误差（0.6）则完全不能令人满意。要将其用于临床实践，还需要更大的数据集。在我们看来，我们的方法论无需修改即可使用。

5 结论

我们提出了一种用于重复空间点模式统计分析的新方法。该方法基于这样一个事实：点过程的概率分布完全由其关联的随机计数测度决定。随机测度可以嵌入再生核希尔伯特空间（RKHS），从而将点过程转化为再生核希尔伯特空间中的随机元素，进而可以应用理论依据充分的方法和算法，这与在欧几里得空间中的做法类似。为此，我们将数据表示在由核的特征函数构成的基底上，并按所需维数截断该表示。这样可以在尽可能减少精度损失的情况下将数据降至较低维数。

作为所提出方法在现实生活中潜在应用的一个例子，我们利用该方法检测了来自三组受试者的人脑锥体神经元位置点模式之间的差异（Diggle 等（ 1991））。我们还利用多个模拟数据集对新观测进行了分类。根据这些实验的结果，可以认为我们的方法适用于实际应用。

参考文献

Aneiros, G., Bongiorno, E.G., Cao, R., Vieu, P., 等人，2017。《函数型统计学及相关领域》。Springer。Aronszajn, N., 1950。再生核理论。《美国数学会汇刊》68，337–404。Baddeley, A., 2015。使用 spatstat 分析重复点模式。Cran 使用指南 35，38。Baddeley, A., B´ar´any, I., Schneider, R., 2007。空间点过程及其应用。《随机几何：2004年9月13–18日在意大利马蒂纳弗兰卡举行的 CIME 暑期学校讲座》，1–75。Barahona, S., Gual-Arnau, X., Ib´a˜nez, M., Sim´o, A., 2018。使用流对儿童身体进行无监督分类。《数据分析与分类进展》12，365–397。Berlinet, A., 1980a。《再生空间与经验测度：函数型估计中的样条方法》。博士论文。Berlinet, A., 1980b。取值于再生核空间的随机变量。《法国科学院院报》290，973–975。Berlinet, A., Thomas-Agnan, C., 2011。《概率与统计中的再生核希尔伯特空间》。Springer Science & Business Media。Carter, D., Prenter, P., 1972。指数空间与计数过程。《概率论与相关领域杂志》21，1–19。Cressie, N., 2015。《空间数据统计学》。约翰·威立父子出版公司。Cuevas, A., 2014。函数型数据统计理论概述。《统计规划与推断杂志》147，1–23。Daley, D., Vere-Jones, D., 1998。随机测度一般理论导论，载于：《点过程理论导论》。Springer，第153–196页。Daley, D., Vere-Jones, D., 2008。随机测度与点过程的基本理论。《点过程理论导论：第二卷：一般理论与结构》，1–75。Diggle, P.J., 2013。《空间和时空点模式的统计分析》。CRC 出版社。Diggle, P.J., Lange, N., Bene˘s, F.M., 1991。临床神经解剖学中重复空间点模式的方差分析。《美国统计学会杂志》86，618–625。

Diggle, P.J., Mateu, J., Clough, H.E., 2000. 参数方法与非参数方法在分析重复空间点模式方面的比较。《应用概率进展》32, 331–343. Eubank, R., Hsing, T., 2008. 随机过程的典型相关。《随机过程及其应用》118, 1634–1661. Ferraty, F., Vieu, P., 2006. 非参数函数型数据分析：理论与实践。Springer Science & Business Media.´Gallego, M.A., Ib´a˜nez, M.V., Sim´o, A., 2016. 芽粒模型的非齐次 k 函数。《空间统计》18, 489–504. Goia, A., Vieu, P., 2016. 高维/无限维统计学最新进展简介。《多元分析杂志》146, 1–6. Gonz´alez, J., Mu˜noz, A., 2010. 再生核希尔伯特空间中的函数型数据表示及其在聚类和分类中的应用。技术报告。马德里卡洛斯三世大学统计系。 Gonz´alez Monsalve, J.A., 等， 2018. 空间和时空点模式比较的统计检验。博士论文。海梅一世大学。 Guilbart, C., 1979. 测度空间上的内积，载于：《IHP概率与统计年刊》，第333–354页。 Hastie, T., Tibshirani, R., Friedman, J.H., 2020. 统计学习要素：数据挖掘、推断与预测。 Horv´ath, L., Kokoszka, P., 2012. 函数型数据推断及其应用。第200卷。Springer Science & Business Media. Hsing, T., Eubank, R., 2015. 函数型数据分析的理论基础：线性算子导论。约翰·威立父子出版公司。 Illian, J., Penttinen, A., Stoyan, H., Stoyan, D., 2008. 空间点模式的统计分析与建模。约翰·威立父子出版公司。 Illian, J.B., Hendrichsen, D.K., 2010. 具有混合效应的吉布斯点过程模型。《环境计量学：国际环境计量学会官方期刊》21, 341–353. Illian, J.B., Sørbye, S.H., Rue, H., 2012. 使用积分嵌套拉普拉斯近似（INLA）拟合复杂空间点过程模型的工具箱。《应用统计年刊》6, 1499–1530. Kadri, H., Duflos, E., Preux, P., Canu, S., Rakotomamonjy, A., Audiffren, J., 2016. 用于函数型响应数据学习的算子值核。《机器学习研究杂志》17, 613–666.

Kupresanin, A., Shin, H., King, D., Eubank, R., 2010. 函数型数据分析的再生核希尔伯特空间框架。《统计规划与推断杂志》140, 3627–3637. Ledoux, M., Talagrand, M., 1991. 巴拿赫空间中的概率：等周问题与随机过程。第23卷。Springer Science & Business Media. Luki´c, M., Beder, J., 2001. 样本路径取值于再生核希尔伯特空间的随机过程。《美国数学会会刊》353, 3945–3969. Mateu, J., Schoenberg, F.P., Diez, D.M., Gonz´alez, J.A., Lu, W., 2015. 点模式间差异度量：基于原型和多维尺度分析的分类。《生物计量学杂志》57, 340–358. Pawlasov´a, K., Dvoˇr´ak, J., 2022. 重复点模式背景下的监督式非参数分类。《图像分析与立体学》41, 57–109. Preda, C., 2007. 基于再生核希尔伯特空间方法的函数型数据回归模型。《统计规划与推断杂志》137, 829–840. R 核心团队，2021. R：统计计算语言与环境。R 统计计算基金会。维也纳，奥地利。网址：https:// www.R-project.org/. Ram´on, P., de la Cruz, M., Chac´on-Labella, J., Escudero, A., 2016. 一种用于生态学中重复点模式分析的新型非参数方法。《生态地理学》39, 1109–1117. Saitoh, S., Sawano, Y., 2016. 再生核理论及其应用。Springer. Sch¨olkopf, B., Smola, A.J., Bach, F., 等，2002. 核方法学习：支持向量机、正则化、优化及其他。麻省理工学院出版社。 Silverman, B., Ramsay, J., 2005. 函数型数据分析。Springer. Smale, S., Zhou, D.X., 2009. 概率空间上的几何。《构造性逼近》30, 311–323. Stoyan, D., Kendall, W., Mecke, J., 1988. 随机几何及其应用。《美国数学会通报》19, 520–523. Stoyan, D., Stoyan, H., 1994. 分形、随机形状与点场：几何统计学方法。约翰·威立父子出版公司。 Suquet, C., 1986. 自再生空间与随机测度。博士论文。里尔第一大学。 Venables, W.N., Ripley, B.D., 2013. 使用 S-PLUS 的现代应用统计学。Springer Science & Business Media.